Homework 1

Zichen Pan zp2197

Problem 1

(a) $L(\lambda) = \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!} e^{-N\lambda}$

(b) $l(\lambda) = \ln L(\lambda) = \ln \lambda \cdot \sum_{i=1}^{N} x_i - N\lambda - \ln \prod_{i=1}^{N} x_i!$

$$\frac{\partial l(\lambda)}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{N} x_i - N \equiv 0 \qquad \lambda_{ML} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$$

(c) $\lambda_{MAP} = \arg\max_{\lambda} \ln p(\lambda | X) = \arg\max_{\lambda} \ln \frac{p(X|\lambda) \cdot p(\lambda)}{p(X)}$

$$p(\lambda) = \text{gamma}(a, b) = \frac{b^a \cdot \lambda^{a-1} \cdot e^{-b\lambda}}{\Gamma(a)}$$

$$\lambda_{MAP} = \arg\max_{\lambda} \sum_{i=1}^{N} x_i \cdot \ln\lambda - N\lambda - \ln \prod_{i=1}^{N} x_i! + a\ln b + (a-1)\ln\lambda - b\lambda - \ln\Gamma(a)$$

let $\lambda_{MAP} = \arg\max_{\lambda} K$.

$$\frac{\partial K}{\partial \lambda} = \frac{\sum_{i=1}^{N} x_i}{\lambda} - N + \frac{a-1}{\lambda} - b \equiv 0 \qquad \lambda_{MAP} = \frac{\sum_{i=1}^{N} x_i + a - 1}{b + N}$$

(d) $p(\lambda | X) = \frac{p(X|\lambda) \cdot p(\lambda)}{p(X)} = \frac{p(X|\lambda) \cdot p(\lambda)}{\int_0^\infty p(X|\lambda) \cdot p(\lambda)\, d\lambda}$

$$= \frac{\frac{\lambda^{\sum_{i=1}^{N} x_i}}{\cancel{\prod_{i=1}^{N} x_i!}} e^{-N\lambda} \cdot \frac{\cancel{b^a} \cdot \lambda^{a-1} \cdot e^{-b\lambda}}{\cancel{\Gamma(a)}}}{\cancel{\frac{b^a}{\prod_{i=1}^{N} x_i! \cdot \Gamma(a)}} \int_0^\infty \lambda^{\sum_{i=1}^{N} x_i} \cdot e^{-N\lambda} \cdot \lambda^{a-1} \cdot e^{-b\lambda}\, d\lambda} = \frac{\lambda^{\sum_{i=1}^{N} x_i + a - 1} \cdot e^{-(N+b)\lambda}}{\int_0^\infty \lambda^{\sum_{i=1}^{N} x_i + a - 1} \cdot e^{-(N+b)\lambda}\, d\lambda}$$

$$\int_0^\infty \lambda^A \cdot e^{-B\lambda}\, d\lambda = \int_0^\infty \lambda^A \cdot \left(-\frac{1}{B} e^{-B\lambda}\right)' d\lambda = \lambda^A \left(-\frac{1}{B} e^{-B\lambda}\right)\Big|_0^\infty + \int_0^\infty A\lambda^{A-1} \cdot \frac{1}{B} e^{-B\lambda}\, d\lambda$$

$$= \int_0^\infty \frac{A}{B} \lambda^{A-1} \cdot e^{-B\lambda}\, d\lambda = \frac{A}{B} \cdot \frac{A-1}{B} \int_0^\infty \lambda^{A-2} \cdot e^{-B\lambda}\, d\lambda = \frac{A!}{B^A} \int_0^\infty e^{-B\lambda}\, d\lambda$$

$$= \frac{A!}{B^A} \cdot \left(-\frac{1}{B} e^{-B\lambda}\right)\Big|_0^\infty = \frac{A!}{B^{A+1}}$$

let $A = \sum_{i=1}^{N} x_i + a - 1$, $B = N + b$

$$\Rightarrow p(\lambda | X) = \frac{\lambda^{\sum_{i=1}^{N} x_i + a - 1} \cdot e^{-(N+b)\lambda} \cdot (N+b)^{1 + \sum_{i=1}^{N} x_i + a - 1}}{\Gamma(1 + \sum_{i=1}^{N} x_i + a - 1)}$$

$$= \text{gamma}\left(\sum_{i=1}^{N} x_i + a, N + b\right)$$

$$(e)\quad p(\lambda|X) = \text{gamma}\left(\sum_{i=1}^{N} x_i + a,\ N+b\right)$$

If $X = \text{gamma}(A, B)$,

$$\phi(t) = \left(\frac{B}{B-t}\right)^A \qquad \phi'(t) = \frac{AB^a}{(B-t)^{A+1}} \qquad \phi''(t) = \frac{A(A+1)B^A}{(B-t)^{A+2}}$$

$$E[X] = \phi'(0) = \frac{A}{B}$$

$$Var(x) = E[x^2] - E^2[x] = \phi''(0) - \left(\frac{A}{B}\right)^2 = \frac{A}{B^2}$$

$$\Rightarrow E[\lambda|X] = \frac{\sum_{i=1}^{N} x_i + a}{N+b}, \quad Var(\lambda|X) = \frac{\sum_{i=1}^{N} x_i + a}{(N+b)^2}$$

The mean of λ under this posterior is more similar with λ_{MAP} because they both take the prior into consideration. λ_{MAP} is the point where $p(\lambda|x)$ is maximal, that's why it is slightly different from the mean. As to $\lambda_{ML,}$ it does not take the prior into the consideration so it lacks some prior distribution parameters in its expression. But basically the format of expression is also similar to the mean.

Zichen Pan zp2197

Problem 2

Assumption: $y \sim N(Xw, \sigma^2 I)$, $\mu = Xw$, $\Sigma = \sigma^2 I$.

$E[w_{RR}] = E[(\lambda I + X^TX)^{-1}X^Ty] = (\lambda I + X^TX)^{-1}X^T \cdot E[y]$

$= (\lambda I + X^TX)^{-1}X^T \cdot Xw$

$Var(y) = E[yy^T] - E[y]E[y^T] = E[yy^T] - \mu\mu^T = \Sigma$

$\Rightarrow E[yy^T] = \mu\mu^T + \Sigma$

$Var(w_{RR}) = E[w_{RR}w_{RR}^T] - E[w_{RR}]E[w_{RR}]^T$

$= E[(\lambda I + X^TX)^{-1}X^T \cancel{Xw}\, y y^T X(\lambda I + X^TX)^{-1}]$

$- (\lambda I + X^TX)^{-1}X^TXw \cdot w^TX^TX(\lambda I+X^TX)^{-1}$

$= (\lambda I + X^TX)^{-1}X^T \cdot E[yy^T]\cdot X(\lambda I + X^TX)^{-1}$

$-(\lambda I + X^TX)^{-1}X^TXww^TX^TX(\lambda I + X^TX)^{-1}$

$= (\lambda I + X^TX)^{-1}X^T\cdot Xww^TX^T\cdot X(\lambda I + X^TX)^{-1}$

$+ (\lambda I + X^TX)^{-1}X^T\cdot\sigma^2 I\cdot X(\lambda I+X^TX)^{-1}$

$-(\lambda I + X^TX)^{-1}X^TXww^TX^TX(\lambda I+X^TX)^{-1}$

$= \sigma^2(\lambda I+X^TX)^{-1}\cdot X^TX(\lambda I+X^TX)^{-1}$

$= \sigma^2[X^TX(\lambda(X^TX)^{-1}+I)]^{-1}\cdot X^TX\cdot \cancel{[X^TX(\lambda(X^TX)^{-1}+I)]^{-1}}\, [(\lambda(X^TX)^{-1}+I)\cdot X^TX]^{-1}$

$= \sigma^2(\lambda(X^TX)^{-1}+I)^{-1}\cancel{(X^TX)^{-1}\cdot X^TX}\cdot(X^TX)^{-1}\cdot(\lambda(X^TX)^{-1}+I)^{-1}$

let $Z = (\lambda(X^TX)^{-1}+I)^{-1}$ $\sigma^2 Z(X^TX)^{-1}Z^T$

Additional proof: $Z = Z^T$

$Z^T = [(\lambda(X^TX)^{-1}+I)^{-1}]^T = [(\lambda(X^TX)^{-1}+I)^T]^{-1} = (\lambda[(X^TX)^{-1}]^T+I)^{-1}$

$= (\lambda[(X^TX)^T]^{-1}+I)^{-1} = (\lambda(X^TX)^{-1}+I)^{-1} = Z.$

Problem 3

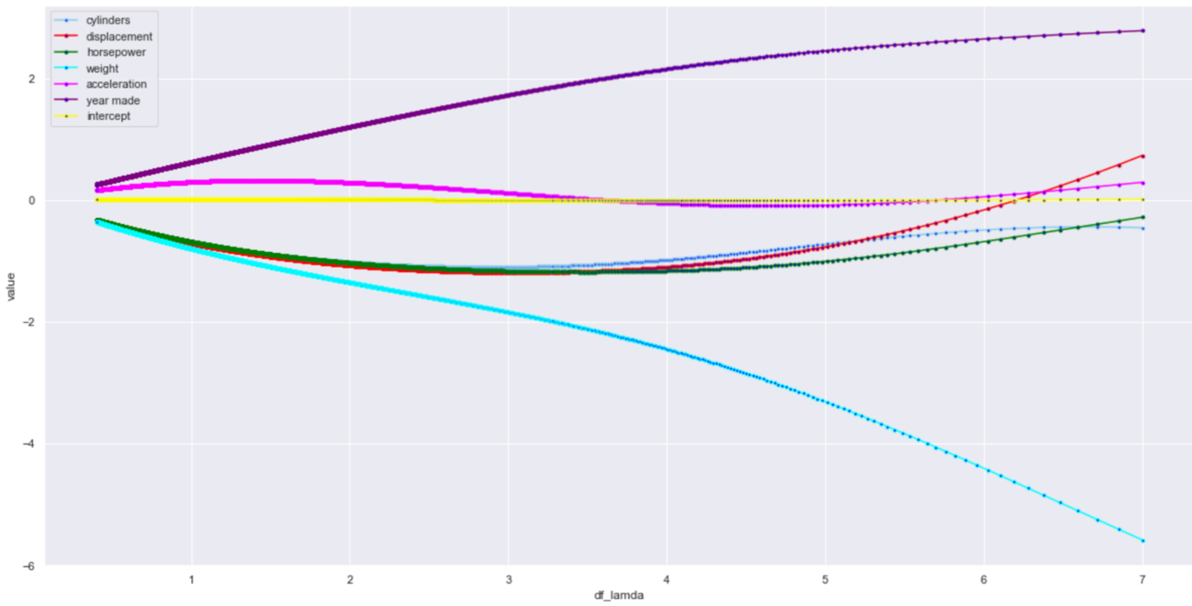
Part 1

(a)

w_{RR} .head(): (df_lamda column included)

	cylinders	displacement	horsepower	weight	acceleration	year made	intercept	df_lamda
0	-0.456261	0.730167	-0.284619	-5.585589	0.289578	2.781398	0.010157	7.000000
1	-0.445724	0.577767	-0.344497	-5.409686	0.251106	2.763335	0.008127	6.850483
2	-0.441310	0.445740	-0.399178	-5.250289	0.216905	2.746405	0.006363	6.715540
3	-0.441428	0.330217	-0.449200	-5.104980	0.186371	2.730449	0.004816	6.592778
4	-0.444919	0.228253	-0.495043	-4.971818	0.159010	2.715338	0.003450	6.480326

plot:

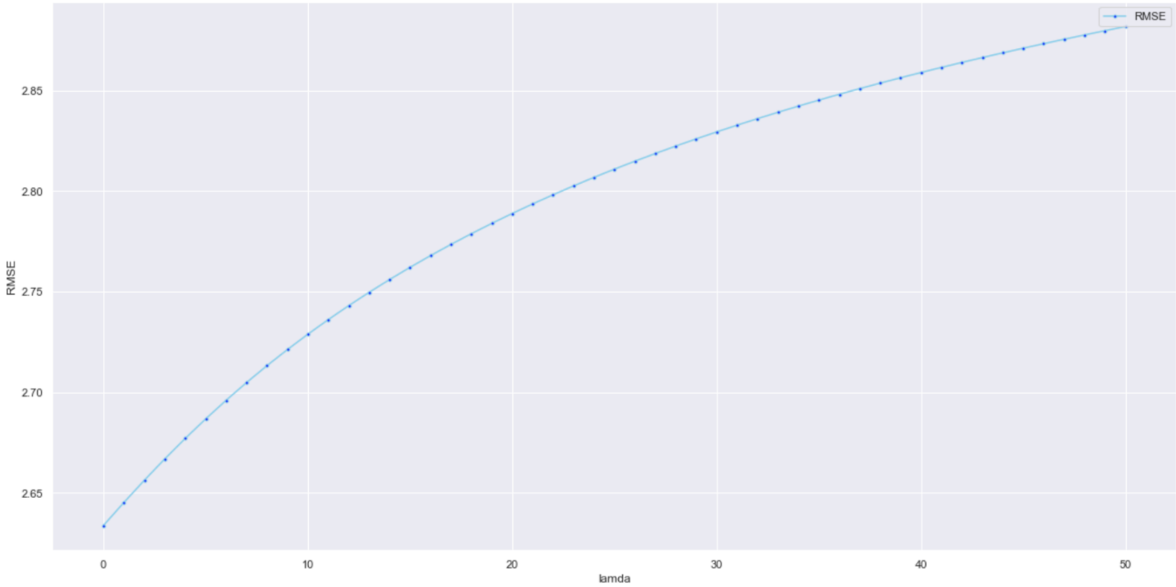


(b)

Two dimensions clearly stand out over the others are weight and year made. They are sensitive to the change of λ, which means their coefficients can be large in linear regression without penalty term in loss function.

(c)

plot:



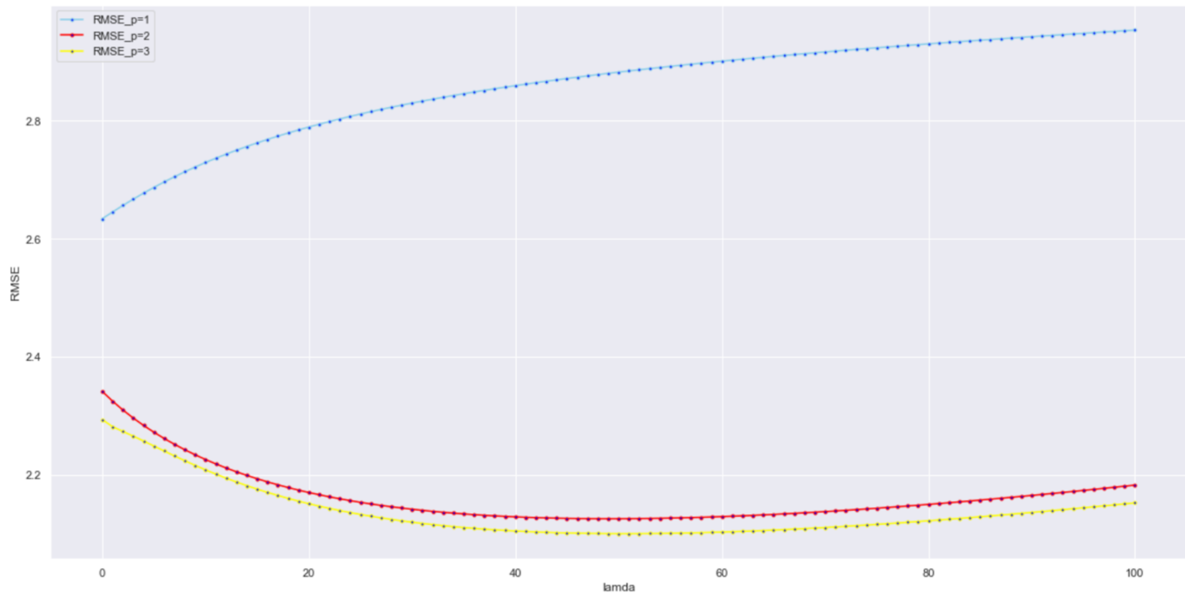
We choose λ=0 for this problem.

When the number of input features is large or the input features are highly correlated, linear regression (least squares) has a tendency to overfit and is not a good option.

Part 2

(d)

plot:



We choose p=3, because at any point the curve of p=3 always has the lowest RMSE on test data.
For λ, we choose λ=50 to reach the minimum of RMSE on test data.